Approximate joint diagonalization with Riemannian optimization on the general linear group

Florent Bouchard, Bijan Afsari, Jérôme Malick, Marco Congedo

APPROXIMATE JOINT DIAGONALIZATION WITH RIEMANNIAN OPTIMIZATION ON THE GENERAL LINEAR GROUP*

FLORENT BOUCHARD†, BIJAN AFSARI , JÉRÔME MALICK‡, AND MARCO CONGEDO†

Abstract. We consider the classical problem of approximate joint diagonalization of matrices, which can be cast as an optimization problem on the general linear group. We propose a versatile Riemannian optimization framework for solving this problem – unifiying existing methods and creating new ones. We use two standard Riemannian metrics (left- and right-invariant metrics) having opposite features regarding the structure of solutions and the model. We introduce the Riemannian optimization tools (gradient, retraction, vector transport) in this context, for the two standard non-degeneracy constraints (oblique and non-holonomic constraints). We also develop tools beyond the classical Riemannian optimization framework to handle the non-Riemannian quotient manifold induced by the non-holonomic constraint with the right-invariant metric. We illustrate our theoretical developments with numerical experiments on both simulated data and a real electroencephalographic recording.

Key words. approximate joint diagonalization, Riemannian optimization, oblique constraint, non-holonomic constraint.

AMS subject classifications. 15A23, 49M15, 58B20, 65K05, 90C25, 90C26, 90C30

1. Introduction.

Approximate joint diagonalization (AJD) is used to solve the well known *blind source separation* (BSS) problem, with applications in a wide variety of engineering fields such as communications, image processing, audio and biomedical signals analysis (see the reference book [16]). In these applications we observe the set $\{C_k\}$ of K $n \times n$ symmetric matrices and we assume they are generated under model

$$C_k = A\Lambda_k A^T + N_k, \quad 1 \leq k \leq K,$$

where $A \in \mathrm{GL}_n$ (group of $n \times n$ invertible matrices) is the mixing matrix, Λ_k are diagonal matrices corresponding to the statistics of the source process and N_k comprise estimation error and measurement noise. AJD aims at finding an approximate joint diagonalizer $B \in \mathbb{R}^{n \times n}$ of the matrices C_k defined as the matrix that minimizes a criterion f measuring the degree of non-diagonality of the set $\{BC_kB^T\}$. In early studies, *e.g.*, [13, 14, 21], the joint diagonalizer B was assumed orthogonal after a prewhitening step meant to orthogonalize the data. However, it is well known that this induces irreversible errors and research has turned toward methods seeking B in GL_n. Thus, AJD can be expressed as an optimization problem over GL_n

$$\inf_{B \in \mathrm{GL}_n} \quad f(B). \tag{1}$$

We refer to [7, 11] for recent studies on suitable AJD criteria and to [5] for the identifiability conditions. There are no analytical solutions to (1) for all standard AJD cost functions. An iterative optimization process over GL_n is needed in general and many algorithms have been proposed in previous studies, see *e.g.*, [1, 4–7, 11, 30, 32–34, 36, 38].

*This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01), funded by the French program "Investissement d'avenir", and the European Research Council, project CHESS 2012-ERC-AdG-320684.

†Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France (florent.bouchard@gipsa-lab.fr).

‡Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

The solution $B \in \mathrm{GL}_n$ is not unique: given any matrices $P \in \mathbb{P}_n$ (group of $n \times n$ permutation matrices) and $\Sigma \in \mathcal{D}_n^*$ (group of $n \times n$ non-singular diagonal matrices), $P\Sigma B$ is a solution equivalent to B, therefore the set of AJD solutions is the whole equivalence class

$$\{P\Sigma B : P \in \mathbb{P}_n, \Sigma \in \mathcal{D}_n^*\}. \tag{2}$$

The permutation ambiguity is of no concern in practice, however the diagonal scaling ambiguity strongly impacts the solution process. Since we can construct sequences (Σ_i) of matrices in $\mathcal{D}_n^*$ converging toward singular matrices, every equivalence class (2) has limit points corresponding to degenerate solutions that must be avoided in practice. Hence, additional constraints on B are needed and various possibilities have been considered in the litterature. A first choice is to fix the *norm of the rows* of B, as for example with the *oblique* constraint [1, 11, 34]. This constraint is expressed as:

$$\mathrm{ddiag}(BB^T) = I_n \qquad \text{(oblique constraint)}, \tag{3}$$

where $\mathrm{ddiag}(\cdot)$ cancels the off-diagonal elements of its argument and I_n denotes the identity matrix. Another constraint that exploits the geometry of the equivalence class is the *non-holonmic* constraint, introduced for BSS in [9] (see in particular [9, section 4] for a perspective on the terminology "non-holonomic") and used for AJD in [5,6,38]. This constraint features the following equivalence class:

$$\{\Sigma B : \Sigma \in \mathcal{D}_n^*\} \qquad \text{(non-holonomic constraint class)}. \tag{4}$$

In contrast with the oblique constraint, using the non-holonomic constraint cancels the action of the diagonal scaling in (2).

The goal of this paper is to propose a unified Riemannian optimization framework rich enough to encompass all existing AJD models featuring different constraints and objective functions. Riemannian optimization over GL_n has already been considered in the context of AJD and BSS in [1] with the Euclidean metric and in [5, 6, 8, 9, 36] with the so-called right-invariant metric (which has a natural link with the model of the AJD and BSS problems; see [8, 9] for details). In our previous work [11] we have proposed a first Riemannian framework adapted to AJD and BSS. However, the approach there is indirect as it uses the polar decomposition to construct surrogate manifolds instead of working directly in the natural space GL_n. This indirect approach fails in particular to encompass the non-holonomic constraint.

In this paper, we develop a general and unified Riemannian optimization framework for AJD constructed directly from GL_n. Beyond the fresh viewpoint, the main contributions of this work are the following:

- We propose to use for AJD the existing Riemannian exponential maps associated with the left- and right-invariant metrics on GL_n.
- We exploit for the first time in the context of AJD a left-invariant metric on GL_n, which is particularly attractive as it is invariant along the equivalence classes (2). We also unify existing results for the right-invariant metric.
- We define the Riemannian submanifolds of GL_n embedding the oblique constraint (3) for both the left- and right-invariant metrics.
- We extend the study of the non-holonomic constraint on three different aspects. First, for the left-invariant metric, we introduce a Riemannian quotient manifold of GL_n to embed it. Second, for the right-invariant metric, we propose basic optimization tools to treat the associated non-Riemannian quotient manifold. Finally, we discuss the possibility to optimize criteria that are not invariant by diagonal scaling.

The outline of this paper is as follows: section 2 summarizes some concepts on Riemannian geometry and optimization that we are going to use intensively in the sequel. In section 3 we study the embedding of the oblique constraint (3) in Riemannian submanifolds of GL_n. In section 4 we focus on the non-holonomic constraint, which deserves an original treatment going beyond standard Riemannian techniques. Section 5 contains numerical experiments both on simulated data and real electroencephalographic recordings.

2. Background on Riemannian geometry and AJD. We first recall the notions of Riemannian optimization that we use in order to develop our framework in section 2.1. We then describe GL_n as a Riemannian manifold when endowed with the left- and right-invariant metrics in section 2.2. Finally, in section 2.3 we define the three AJD criteria that we use in our numerical experiments.

2.1. Riemannian optimization in a nutshell. This section contains a brief introduction on the necessary ingredients for gradient based Riemannian optimization on matrix manifolds. We refer to [2] for a complete coverage of this topic.

A smooth matrix manifold $\mathcal{M}$ admits a differentiable structure and every point $x \in \mathcal{M}$ has a *tangent space* $T_x\mathcal{M}$. Such $\mathcal{M}$ is turned into a Riemannian manifold by endowing it with a *Riemannian metric* $\langle \cdot, \cdot \rangle_\cdot$, which is a smoothly varying inner product on every tangent space $T_x\mathcal{M}$. The Riemannian manifold $\mathcal{M}$ equipped with $\langle \cdot, \cdot \rangle_\cdot$ is a curved space for which *geodesics* $\gamma : \mathbb{R} \to \mathcal{M}$ generalize the concept of straight lines. A geodesic γ only depends on the choice of the initial point $\gamma(0) = x \in \mathcal{M}$ and initial direction $\dot\gamma(0) = \xi \in T_x\mathcal{M}$, where $\dot\gamma$ denotes the derivative of γ. It is also possible to translate tangent vectors along a curve in $\mathcal{M}$ so that they remain parallel with respect to the Riemannian metric by using *parallel transport*. Given $\gamma : \mathbb{R} \to \mathcal{M}$ such that $\gamma(0) = x \in \mathcal{M}$ and $\dot\gamma(0) = \xi \in T_x\mathcal{M}$, the parallel transport $t \mapsto P(t)$ along γ with initial value $P(0) = \eta \in T_x\mathcal{M}$ is the vector field such that $P(t)$ is the tangent vector in $T_{\gamma(t)}\mathcal{M}$ parallel to η.

Concerning Riemannian optimization, given an objective function $f : \mathcal{M} \to \mathbb{R}$, the *Riemannian gradient* $\operatorname{grad}_\mathcal{M} f(x)$ of f at $x \in \mathcal{M}$ is defined through the Riemannian metric as the unique tangent vector in $T_x\mathcal{M}$ such that for all $\xi \in T_x\mathcal{M}$

$$\langle \operatorname{grad}_\mathcal{M} f(x), \xi \rangle_x = \mathrm{D}\, f(x)[\xi],$$

where $\mathrm{D}\, f(x)[\xi]$ is the directional derivative of f at x in the direction ξ. A descent direction of a criterion f at $x \in \mathcal{M}$ is a tangent vector $\xi \in T_x\mathcal{M}$ such that

$$\langle \operatorname{grad}_\mathcal{M} f(x), \xi \rangle_x < 0.$$

A new point on the manifold is then achieved by a *retraction* R on $\mathcal{M}$, which is a mapping from the tangent spaces back onto $\mathcal{M}$ satisfying, for all $x \in \mathcal{M}$ and $\xi \in T_x\mathcal{M}$,

$$R_x(0_x) = x \qquad\qquad \mathrm{D}\, R_x(0_x)[\xi] = \xi,$$

where 0_x denotes the zero element of $T_x\mathcal{M}$. Every Riemannian manifold $\mathcal{M}$ equipped with a metric $\langle \cdot, \cdot \rangle_\cdot$ admits a natural retraction that arises from its geodesics: the *Riemannian exponential map*, which is defined for all $x \in \mathcal{M}$ and $\xi \in T_x\mathcal{M}$ as $\exp_x(\xi) = \gamma(1)$, where γ is the geodesic such that $\gamma(0) = x$ and $\dot\gamma(0) = \xi$. The Riemannian exponential map might however be complicated to find or too expensive to compute. In such cases, an alternative retraction is preferable; see [2, 3]. The basic Riemannian gradient descent algorithm, which simply takes $-\operatorname{grad}_\mathcal{M} f(x_i)$ as a descent direction, is

$$x_{i+1} = R_{x_i}(-t_i \operatorname{grad}_\mathcal{M} f(x_i)),$$

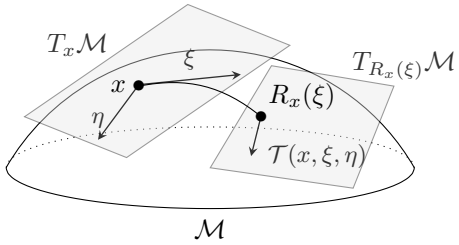

FIG. 1. *Schematic illustration of vector transport on $\mathcal{M}$.*

where t_i is the stepsize, which can for instance be computed with a linesearch; see [2, chapter 4]. Finally, some optimization methods such as conjugate gradient or quasi-Newton use information given by the descent direction of previous iterates in order to obtain the new descent direction. To do so in the Riemannian framework we need to be able to transport a tangent vector of a point in $\mathcal{M}$ into the tangent space of another point. It can be achieved by using a *vector transport* $\mathcal{T}$ on $\mathcal{M}$ associated with a retraction R, closely related to the concept of parallel transport. Given $x \in \mathcal{M}$ and $\xi, \eta \in T_x\mathcal{M}$, the vector transport $\mathcal{T}(x, \xi, \eta)$ transports the vector η in the tangent space of $R_x(\xi)$ as illustrated in figure 1.

2.2. GL_n as a Riemannian manifold. All the results given in this section can be found in [10, 27, 28, 35, 37]. The general linear group GL_n is an open set in $\mathbb{R}^{n\times n}$, thus its tangent space $T_B\mathrm{GL}_n$ at any point B can be identified as $\mathbb{R}^{n\times n}$, which is denoted $T_B\mathrm{GL}_n \simeq \mathbb{R}^{n\times n}$. We endow GL_n either with the left- or right-invariant metric, respectively defined for all $B \in \mathrm{GL}_n$, $\xi, \eta \in T_B\mathrm{GL}_n$ as

$$\langle \xi, \eta \rangle^{\ell}_B = \mathrm{tr}(B^{-1}\xi(B^{-1}\eta)^T) \qquad \text{and} \qquad \langle \xi, \eta \rangle^{r}_B = \mathrm{tr}\left(\xi B^{-1}(\eta B^{-1})^T\right). \tag{5}$$

Complete geodesics, *i.e.*, defined on the whole line $\mathbb{R}$, can be found for both these metrics. The geodesics $\gamma_\ell : \mathbb{R} \to \mathrm{GL}_n$ and $\gamma_r : \mathbb{R} \to \mathrm{GL}_n$ of GL_n equiped with the left- and right-invariant metrics are defined for all $B \in \mathrm{GL}_n$ and $\xi \in T_B\mathrm{GL}_n$ as

$$\gamma_\ell(t) = B \exp\left(t(B^{-1}\xi)^T\right) \exp\left(t\left(B^{-1}\xi - (B^{-1}\xi)^T\right)\right), \tag{6}$$

and

$$\gamma_r(t) = \exp\left(t\left(\xi B^{-1} - (\xi B^{-1})^T\right)\right) \exp\left(t(\xi B^{-1})^T\right) B, \tag{7}$$

where $\exp(\cdot)$ denotes the matrix exponential. The corresponding Riemannian exponential maps at $B \in \mathrm{GL}_n$ are denoted $\exp^{\ell}_B : T_B\mathrm{GL}_n \to \mathrm{GL}_n$ and $\exp^{r}_B : T_B\mathrm{GL}_n \to \mathrm{GL}_n$.

Let $f : \mathrm{GL}_n \to \mathbb{R}$ be an objective function and $\mathrm{grad}_{\mathrm{E}}\, f(B)$ its Euclidean gradient at $B \in \mathrm{GL}_n$. The Riemannian gradients $\mathrm{grad}_\ell\, f(B)$ and $\mathrm{grad}_r\, f(B)$ of f at B in GL_n equiped with the left- and right-invariant metrics are given by

$$\mathrm{grad}_\ell\, f(B) = BB^T\, \mathrm{grad}_{\mathrm{E}}\, f(B) \qquad \text{and} \qquad \mathrm{grad}_r\, f(B) = \mathrm{grad}_{\mathrm{E}}\, f(B) B^T B.$$

Finally, the vector transport on GL_n can simply be defined by $\mathcal{T}(B, \xi, \eta) = \eta$. However, it appears more natural to use the vector transports

$$\mathcal{T}_\ell(B, \xi, \eta) = \exp^{\ell}_B(\xi) B^{-1}\eta \qquad\qquad \mathcal{T}_r(B, \xi, \eta) = \eta B^{-1} \exp^{r}_B(\xi),$$

which preserve the left- and right-invariant metrics, respectively. Notice that the scaling by B^{-1} helps in avoiding the boundary of GL_n. We thus have all necessary

tools for gradient based Riemannian optimization on GL_n equipped with the left- or right-invariant metrics as defined in (5).

2.3. AJD objective functions. In this section, we recall three standard AJD objective functions to be minimized that we use in our numerical experiments along with their Euclidean gradients, which are needed to obtain the Riemannian counterparts. The first objective function that we consider is the classical least squares criterion [13, 14] based on the Frobenius distance, defined as

$$f_{\mathrm{F}}(B) = \sum_k \left\| BC_kB^T - \mathrm{ddiag}(BC_kB^T) \right\|_F^2, \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathrm{ddiag}(\cdot)$ cancel the off-diagonal elements of its argument. Its Euclidean gradient $\mathrm{grad}_{\mathrm{E}}\, f_{\mathrm{F}}(B)$ at $B \in \mathrm{GL}_n$ is given by

$$\mathrm{grad}_{\mathrm{E}}\, f_{\mathrm{F}}(B) = \sum_k 4 \left[BC_kB^T - \mathrm{ddiag}(BC_kB^T) \right] BC_k.$$

This criterion is not invariant to diagonal scaling, *i.e.*, we do not have $f_{\mathrm{F}}(\Sigma B) = f_{\mathrm{F}}(B)$ for all $B \in \mathrm{GL}_n$ and $\Sigma \in \mathcal{D}_n^*$. To overcome this problem, some authors have proposed modified versions of this criterion that are invariant to diagonal scaling of B [5, 7]. In this article we consider the objective function proposed in [5], which is defined as

$$\widetilde{f}_{\mathrm{F}}(B) = \sum_k \left\| C_k - B^{-1}\, \mathrm{ddiag}(BC_kB^T) B^{-T} \right\|_F^2. \tag{9}$$

Its Euclidean gradient $\mathrm{grad}_{\mathrm{E}}\, \widetilde{f}_{\mathrm{F}}(B)$ at $B \in \mathrm{GL}_n$ is given by

$$\mathrm{grad}_{\mathrm{E}}\, \widetilde{f}_{\mathrm{F}}(B) = \sum_k 4 \left[Q_k(B)\, \mathrm{ddiag}(BC_kB^T) - \mathrm{ddiag}(Q_k(B)) BC_kB^T \right] B^{-T},$$

where $Q_k(B) = (BB^T)^{-1}[BC_kB^T - \mathrm{ddiag}(BC_kB^T)](BB^T)^{-1}$. As explained in [5], it is interesting to notice that criteria (8) and (9) are equal if B is orthogonal. An advantage of criterion (9) is that its domain of definition regarding matrices C_k remains the set of symmetric matrices whereas they further have to be positive definite in the modified version proposed in [7]. The last diagonality criterion that we consider is the one based on the log-likelihood [30, 31], defined as

$$f_{\mathrm{ll}}(B) = \sum_k \log \frac{\det(\mathrm{ddiag}(BC_kB^T))}{\det(BC_kB^T)}. \tag{10}$$

Criterion (10) is only defined for symmetric positive definite matrices C_k and possesses the diagonal scaling invariance property. Its Euclidean gradient $\mathrm{grad}_{\mathrm{E}}\, f_{\mathrm{ll}}(B)$ at $B \in \mathrm{GL}_n$ is given by

$$\mathrm{grad}_{\mathrm{E}}\, f_{\mathrm{ll}}(B) = \sum_k 2 \left[\mathrm{ddiag}(BC_kB^T)^{-1} - (BC_kB^T)^{-1} \right] BC_k.$$

3. Oblique constraint. In this section we formalize the necessary tools for gradient-based Riemannian optimization over the submanifold of GL_n resulting from the oblique constraint (3), for both left- and right-invariant metrics defined in (5). The oblique constraint (3) yields the submanifold of GL_n (see [1]) defined as

$$\mathcal{M}_n^{\mathrm{o}} = \{ B \in \mathrm{GL}_n : \mathrm{ddiag}(BB^T) = I_n \},$$

whose tangent space $T_B\mathcal{M}_n^{\mathrm{o}}$ at $B \in \mathcal{M}_n^{\mathrm{o}}$ is

$$T_B\mathcal{M}_n^{\mathrm{o}} = \{\xi \in \mathbb{R}^{n\times n} : \operatorname{ddiag}(\xi B^T) = 0\}.$$

$\mathcal{M}_n^{\mathrm{o}}$ is turned into a Riemannian submanifold of GL_n by inheriting either the left- or right-invariant metric. For all $B \in \mathcal{M}_n^{\mathrm{o}}$, we can define the *orthogonal projection map* from the ambient space $\mathbb{R}^{n\times n}$ onto $T_B\mathcal{M}_n^{\mathrm{o}}$ according to the chosen metric, which is then used to obtain the Riemannian gradient and vector transport. These projection maps are given in propositions 3.1 and 3.2.

PROPOSITION 3.1. *The orthogonal projection map $P_B^{\mathrm{o},\ell} : \mathbb{R}^{n\times n} \to T_B\mathcal{M}_n^{\mathrm{o}}$ at $B \in \mathcal{M}_n^{\mathrm{o}}$ equipped with the left-invariant metric is, for all $Z \in \mathbb{R}^{n\times n}$,*

$$P_B^{\mathrm{o},\ell}(Z) = Z - BB^T\Lambda B, \qquad \operatorname{diag}(\Lambda) = \left(BB^T \odot BB^T\right)^{-1} \operatorname{diag}(ZB^T),$$

where $\Lambda \in \mathcal{D}_n$ (set of diagonal matrices), $\operatorname{diag}(\cdot)$ returns the vector containing the diagonal elements of its argument and $\odot$ denotes the Hadamard product.

Proof. First, we need to determine the normal space to $T_B\mathcal{M}_n^{\mathrm{o}}$ according to the left-invariant metric, which is defined as $\{\eta \in \mathbb{R}^{n\times n} : \langle \xi, \eta\rangle_B^\ell = 0, \forall \xi \in T_B\mathcal{M}_n^{\mathrm{o}}\}$. It is $\{BB^T\Lambda B : \Lambda \in \mathcal{D}_n\}$. Indeed, it has the correct dimension, which is n, and

$$\langle \xi, BB^T\Lambda B\rangle_B^\ell = \operatorname{tr}(\xi B^T\Lambda) = \operatorname{tr}(\operatorname{ddiag}(\xi B^T)\Lambda) = 0.$$

Thus, $P_B^{\mathrm{o},\ell}(Z) = Z - BB^T\Lambda B$ and $\operatorname{ddiag}(P_B^{\mathrm{o},\ell}(Z)B^T) = 0$ can be vectorized as

$$(BB^T \odot BB^T)\operatorname{diag}(\Lambda) = \operatorname{diag}(ZB^T).$$

Since $B \in \mathrm{GL}_n$, BB^T is positive definite and the Schur product theorem ensures that $BB^T \odot BB^T$ is invertible. This completes the proof. ∎

PROPOSITION 3.2. *The orthogonal projection map $P_B^{\mathrm{o},r} : \mathbb{R}^{n\times n} \to T_B\mathcal{M}_n^{\mathrm{o}}$ at $B \in \mathcal{M}_n^{\mathrm{o}}$ equipped with the right-invariant metric is, for all $Z \in \mathbb{R}^{n\times n}$,*

$$P_B^{\mathrm{o},r}(Z) = Z - \operatorname{ddiag}(ZB^T)\operatorname{ddiag}\left((BB^T)^2\right)^{-1} BB^TB.$$

Proof. The normal space to $T_B\mathcal{M}_n^{\mathrm{o}}$ according to the right-invariant metric is $\{\Lambda BB^TB : \Lambda \in \mathcal{D}_n\}$. Indeed, its dimension is n and

$$\langle \xi, \Lambda BB^TB\rangle_B^r = \operatorname{tr}(\xi B^T\Lambda) = \operatorname{tr}(\operatorname{ddiag}(\xi B^T)\Lambda) = 0.$$

Thus, we have $P_B^{\mathrm{o},r}(Z) = Z - \Lambda BB^TB$ and solving $\operatorname{ddiag}(P_B^{\mathrm{o},r}(Z)B^T) = 0$ yields $\Lambda = \operatorname{ddiag}(ZB^T)\operatorname{ddiag}\left((BB^T)^2\right)^{-1}$. ∎

It follows from [2, eq. (3.37)] that the Riemannian gradients $\operatorname{grad}_{\mathrm{o},\ell} f(B)$ and $\operatorname{grad}_{\mathrm{o},r} f(B)$ of an objective function f at $B \in \mathcal{M}_n^{\mathrm{o}}$ endowed with the left- and right-invariant metrics are given by

$$\operatorname{grad}_{\mathrm{o},\ell} f(B) = P_B^{\mathrm{o},\ell}(\operatorname{grad}_\ell f(B)) \qquad \text{and} \qquad \operatorname{grad}_{\mathrm{o},r} f(B) = P_B^{\mathrm{o},r}(\operatorname{grad}_r f(B)).$$

PROPOSITION 3.3. *Given a retraction R on GL_n, a retraction on $\mathcal{M}_n^{\mathrm{o}}$ is defined for all $B \in \mathcal{M}_n^{\mathrm{o}}$ and $\xi \in T_B\mathcal{M}_n^{\mathrm{o}}$ by*

$$R_B^{\mathrm{o}}(\xi) = \operatorname{ddiag}\left(R_B(\xi)R_B(\xi)^T\right)^{-1/2} R_B(\xi).$$

metric	$\langle\cdot,\cdot\rangle^{\ell}$	$\langle\cdot,\cdot\rangle^{r}$
projection map	$P_B^{\mathrm{o},\ell}(Z) = Z - BB^T\Lambda B$, $\operatorname{diag}(\Lambda) = (BB^T \odot BB^T)^{-1}\operatorname{diag}(ZB^T)$	$P_B^{\mathrm{o},r}(Z) = Z - \Lambda BB^T B$, $\Lambda = \operatorname{ddiag}(ZB^T)\operatorname{ddiag}((BB^T)^2)^{-1}$
gradient	$P_B^{\mathrm{o},\ell}(\operatorname{grad}_\ell f(B))$	$P_B^{\mathrm{o},r}(\operatorname{grad}_r f(B))$
retraction	$R_B^{\mathrm{o},\ell}(\xi) = \Lambda(\exp_B^\ell(\xi))\exp_B^\ell(\xi)$	$R_B^{\mathrm{o},r}(\xi) = \Lambda(\exp_B^r(\xi))\exp_B^r(\xi)$
	where $\Lambda(X) = \operatorname{ddiag}(XX^T)^{-1/2}$	
vector transport	$P^{\mathrm{o},\ell}_{R_B^{\mathrm{o},\ell}(\xi)}(R_B^{\mathrm{o},\ell}(\xi)B^{-1}\eta)$	$P^{\mathrm{o},r}_{R_B^{\mathrm{o},r}(\xi)}(\eta B^{-1}R_B^{\mathrm{o},r}(\xi))$

TABLE 1
Summary of the necessary tools for Riemannian optimization on the oblique manifold $\mathcal{M}_n^{\mathrm{o}}$.

Proof. We need to show that $R_B^{\mathrm{o}}(\xi) = B + \xi + o(\|\xi\|)$. Since $R_B(\xi) = B + \xi + o(\|\xi\|)$, we have

$$\operatorname{ddiag}(R_B(\xi)R_B(\xi)^T) = \operatorname{ddiag}(BB^T) + 2\operatorname{ddiag}(\xi B^T) + o(\|\xi\|).$$

From $\operatorname{ddiag}(BB^T) = I_n$ and $\operatorname{ddiag}(\xi B^T) = 0$, we obtain

$$\operatorname{ddiag}(R_B(\xi)R_B(\xi)^T) = I_n + o(\|\xi\|).$$

Using this and $R_B(\xi) = B + \xi + o(\|\xi\|)$ completes the proof. ∎

Note that we choose to define the retraction R on GL_n of proposition 3.3 as *(i)* the exponential map $\exp^\ell$ resulting from (6) when we endow $\mathcal{M}_n^{\mathrm{o}}$ with the left-invariant metric, and as *(ii)* $\exp^r$ resulting from (7) when endowing it with the right-invariant one. The resulting retractions are denoted $R^{\mathrm{o},\ell}$ and $R^{\mathrm{o},r}$ respectively. Equation (8.10) in [2] shows that the vector transports $\mathcal{T}_{\mathrm{o},\ell}$ and $\mathcal{T}_{\mathrm{o},r}$ on $\mathcal{M}_n^{\mathrm{o}}$ equipped with the left- and right-invariant metrics can be defined for all $B \in \mathcal{M}_n^{\mathrm{o}}$, $\xi, \eta \in T_B\mathcal{M}_n^{\mathrm{o}}$ as

$$\mathcal{T}_{\mathrm{o},\ell}(B,\xi,\eta) = P^{\mathrm{o},\ell}_{R_B^{\mathrm{o},\ell}(\xi)}(R_B^{\mathrm{o},\ell}(\xi)B^{-1}\eta) \text{ and } \mathcal{T}_{\mathrm{o},r}(B,\xi,\eta) = P^{\mathrm{o},r}_{R_B^{\mathrm{o},r}(\xi)}(\eta B^{-1}R_B^{\mathrm{o},r}(\xi)).$$

To conclude this section, a summary of all the necessary tools for Riemannian optimization on the oblique manifold equipped with the left- and right-invariant metrics is given in table 1. The extension to second-order Riemannian optimization methods is straightforward: to define the Riemannian Hessian we need the Levi-Civita connection, which, in the case of the oblique manifold, is obtained by projecting the Levi-Civita connection of GL_n (given for instance in [10, 35, 37]) onto the tangent space; see [2, proposition 5.3.2].

4. Non-holonomic constraint. In this section we focus on embedding the non-holonomic constraint in a quotient manifold of GL_n, denoted $\mathcal{M}_n^{\mathrm{nh}}$. In section 4.1, we specify standard geometrical objects in the quotient manifold and how this inherits a Riemannian structure from an invariant Riemannian structure of GL_n. In section 4.2, we apply these results when GL_n is endowed with the left-invariant metric, yielding a proper Riemannian quotient manifold. In section 4.3, we discuss how we can still optimize on the non-Riemannian quotient manifold $\mathcal{M}_n^{\mathrm{nh}}$ when GL_n is endowed with the right-invariant metric. Finally, in section 4.4, we mention that we can exploit the geometry of $\mathcal{M}_n^{\mathrm{nh}}$ to minimize criteria that do not induce proper functions on $\mathcal{M}_n^{\mathrm{nh}}$.

4.1. Riemannian quotient manifold structure. The non-holonomic constraint yields $\mathcal{M}_n^{\mathrm{nh}} = \mathrm{GL}_n/\mathcal{D}_n^*$ defined as

$$\mathcal{M}_n^{\mathrm{nh}} = \{\{\Sigma B : \Sigma \in \mathcal{D}_n^*\} : B \in \mathrm{GL}_n\},$$

which is a smooth quotient manifold of dimension n^2-n [26, theorem 7.10]. To handle the elements of this manifold, which are equivalence classes, the usual technique [2] is to use the canonical projection π mapping the elements of GL_n onto $\mathcal{M}_n^{\mathrm{nh}}$. Equivalence classes are obtained through π as $\{\Sigma B : \Sigma \in \mathcal{D}_n^*\} = \pi^{-1}(\pi(B))$ and each element $\mathcal{B} \in \mathcal{M}_n^{\mathrm{nh}}$ can be represented by any matrix $B \in \mathrm{GL}_n$ such that $\mathcal{B} = \pi(B)$. The geometry and objects required for optimization on $\mathcal{M}_n^{\mathrm{nh}}$ can be characterized by representations at the matrix level. Note that from a numerical perspective it may be necessary to select a particular representative (*i.e.*, impose a diagonal scaling) in order to avoid degenerate elements of each class.

The tangent space $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{nh}}$ of $\mathcal{B} = \pi(B) \in \mathcal{M}_n^{\mathrm{nh}}$ can be fully described by a subspace of the tangent space of B. In particular, $T_B\mathrm{GL}_n \simeq \mathbb{R}^{n\times n}$ can be decomposed into two complementary subspaces: the *vertical* and *horizontal* spaces [2]. The vertical space is defined as the tangent space $T_B\pi^{-1}(\pi(B))$ of the equivalence class $\pi^{-1}(\pi(B))$ at B. This subspace contains all the elements of $T_B\mathrm{GL}_n$ inducing a displacement along $\pi^{-1}(\pi(B))$. In the case of $\mathcal{M}_n^{\mathrm{nh}}$, the vertical space at $B \in \mathrm{GL}_n$ is

$$\mathcal{V}_B = \{\Delta B : \Delta \in \mathcal{D}_n\}.$$

The horizontal space $\mathcal{H}_B$ at B is then defined as the orthogonal complement to $\mathcal{V}_B$ in $T_B\mathrm{GL}_n$ according to the chosen metric $\langle\cdot,\cdot\rangle_\cdot$ on GL_n. The elements of $\mathcal{H}_B$, called *horizontal lifts*, provide proper representations of the tangent vectors in $T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{nh}}$. Indeed, for all $\Xi \in T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{nh}}$, there is a unique $\xi \in \mathcal{H}_B$ such that $\mathrm{D}\,\pi(B)[\xi] = \Xi$.

The horizontal space thus depends on the choice of the metric $\langle\cdot,\cdot\rangle_\cdot$ on GL_n. For $\mathcal{M}_n^{\mathrm{nh}}$ to be a proper Riemannian quotient manifold, this metric must induce a well-defined Riemannian metric on the quotient. To do so, the metric $\langle\cdot,\cdot\rangle_\cdot$ has to be invariant along each equivalence class. In our case, we must have, for all $B \in \mathrm{GL}_n$, $\Sigma \in \mathcal{D}_n^*$ and $\xi, \eta \in \mathbb{R}^{n\times n}$

$$\langle \Sigma\xi, \Sigma\eta\rangle_{\Sigma B} = \langle \xi, \eta\rangle_B. \tag{11}$$

Notice that the left-invariant metric satisfies (11) but the right-invariant one does not.

Geodesics (and thus Riemannian exponential maps) on $\mathcal{M}_n^{\mathrm{nh}}$ can be obtained through geodesics on GL_n. Indeed, if $\gamma(t)$ is a complete geodesic that stays horizontal in GL_n, *i.e.*, if its derivative $\dot\gamma(t)$ is in $\mathcal{H}_{\gamma(t)}$, then $\pi(\gamma(t))$ is a complete geodesic on $\mathcal{M}_n^{\mathrm{nh}}$ [22, proposition 2.109]. Furthermore, from proposition 4.1.3 in [2], a retraction R on GL_n also induces a retraction on $\mathcal{M}_n^{\mathrm{nh}}$ if we have, for all $B \in \mathrm{GL}_n$, $\xi \in \mathcal{H}_B$ and $\Sigma \in \mathcal{D}_n^*$,

$$\pi(R_B(\xi)) = \pi(R_{\Sigma B}(\Sigma\xi)). \tag{12}$$

Concerning the vector transport, let P be the orthogonal projection map on the horizontal space, R be a retraction on GL_n satisfying (12) and $\mathcal{T}$ be a vector transport on GL_n associated with R. Given $\mathcal{B} = \pi(B) \in \mathcal{M}_n^{\mathrm{nh}}$, $\Xi, H \in T_{\mathcal{B}}\mathcal{M}_n^{\mathrm{nh}}$ with horizontal lifts $\xi, \eta \in \mathcal{H}_B$, a suitable vector transport $\mathcal{T}_{\mathrm{nh}}(\mathcal{B}, \Xi, H)$ on $\mathcal{M}_n^{\mathrm{nh}}$ can be represented by $P_{R_B(\xi)}(\mathcal{T}(B, \xi, \eta))$, *i.e.*,

$$\mathcal{T}_{\mathrm{nh}}(\mathcal{B}, \Xi, H) = \mathrm{D}\,\pi(R_B(\xi))[P_{R_B(\xi)}(\mathcal{T}(B, \xi, \eta))].$$

Finally, a criterion f defined on GL_n induces a well-defined criterion $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$ if it is invariant along each equivalence class, *i.e.*, if $f(\Sigma B) = f(B)$ for all $B \in \mathrm{GL}_n$ and

$\Sigma \in \mathcal{D}_n^*$. In this case, $\widehat{f}$ is the function such that $f = \widehat{f} \circ \pi$ and the optimization of f can be done by optimizing $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$. The representation of the Riemannian gradient of $\widehat{f}$ at $\mathcal{B} = \pi(B) \in \mathcal{M}_n^{\mathrm{nh}}$ is simply the Riemannian gradient of f at B, which belongs to $\mathcal{H}_B$ [2, eq. (3.39)].

4.2. Left-invariant metric. The left-invariant metric in (5) satisfies (11) and therefore induces a proper Riemannian metric on $\mathcal{M}_n^{\mathrm{nh}}$, which becomes a proper Riemannian quotient manifold. To perform gradient-based Riemannian optimization, it remains to define the horizontal space, the orthogonal projection map, a retraction and a vector transport. The horizontal space $\mathcal{H}_B^\ell$ and orthogonal projection map $P_B^{\mathrm{nh},\ell}$ are given in proposition 4.1. Concerning the retraction, we can use the Riemannian exponential map, simply represented by $\exp_B^\ell$, associated to the the geodesics on $\mathcal{M}_n^{\mathrm{nh}}$, which are given in proposition 4.2. Finally, given horizontal lifts $\xi, \eta \in \mathcal{H}_B^\ell$, the vector transport on $\mathcal{M}_n^{\mathrm{nh}}$ can be represented by $P^{\mathrm{nh},\ell}_{\exp_B^\ell(\xi)}(\exp_B^\ell(\xi)B^{-1}\eta)$.

PROPOSITION 4.1. *The horizontal space $\mathcal{H}_B^\ell$ at B in GL_n endowed with the left-invariant metric is*

$$\mathcal{H}_B^\ell = \{\xi \in \mathbb{R}^{n\times n} : \operatorname{ddiag}\left((BB^T)^{-1}\xi B^T\right) = 0\}.$$

It follows that the orthogonal projection map at B from $\mathbb{R}^{n\times n}$ onto $\mathcal{H}_B^\ell$ is

$$P_B^{\mathrm{nh},\ell}(Z) = Z - \Lambda B, \qquad \operatorname{diag}(\Lambda) = \left((BB^T)^{-1} \odot BB^T\right)^{-1} \operatorname{diag}\left((BB^T)^{-1}ZB^T\right),$$

where $\Lambda \in \mathcal{D}_n$.

Proof. The set $\mathcal{H}_B^\ell$ is of dimension $n^2 - n$ and for all $B \in \mathrm{GL}_n$, $\Delta \in \mathcal{D}_n$ and $\xi \in \mathbb{R}^{n\times n}$, we have

$$\langle \xi, \Delta B\rangle_B^\ell = \operatorname{tr}((BB^T)^{-1}\xi B^T\Delta) = \operatorname{tr}(\operatorname{ddiag}((BB^T)^{-1}\xi B^T)\Delta).$$

Thus, $\langle \xi, \Delta B\rangle_B^\ell = 0$ for all $\Delta \in \mathcal{D}_n$ if and only if $\operatorname{ddiag}((BB^T)^{-1}\xi B^T) = 0$. For the orthogonal projection map, we know that $P_B^{\mathrm{nh},\ell}(Z) = Z - \Lambda B$ and equation $\operatorname{ddiag}((BB^T)^{-1}P_B^{\mathrm{nh},\ell}(Z)B^T) = 0$ can be vectorized as

$$\left((BB^T)^{-1} \odot BB^T\right)\operatorname{diag}(\Lambda) = \operatorname{diag}\left((BB^T)^{-1}ZB^T\right).$$

Since $B \in \mathrm{GL}_n$, both BB^T and $(BB^T)^{-1}$ are positive definite and the Schur product theorem ensures that $(BB^T)^{-1} \odot BB^T$ is invertible. This is enough to conclude. ∎

PROPOSITION 4.2. *For all $B \in \mathrm{GL}_n$ and $\xi \in \mathcal{H}_B^\ell$, $\gamma_{\mathrm{nh},\ell} : \mathbb{R} \to \mathcal{M}_n^{\mathrm{nh}}$ defined as*

$$\gamma_{\mathrm{nh},\ell}(t) = \pi(\gamma_\ell(t)),$$

where γ_ℓ is defined in (6), are complete geodesics on $\mathcal{M}_n^{\mathrm{nh}}$.

Proof. The derivative of $\gamma_\ell(t)$ is given by

$$\dot\gamma_\ell(t) = B\exp\left(t(B^{-1}\xi)^T\right)B^{-1}\xi\exp\left(t\left(B^{-1}\xi - (B^{-1}\xi)^T\right)\right),$$

and it follows that for all $t \in \mathbb{R}$

$$\operatorname{ddiag}\left((\gamma_\ell(t)\gamma_\ell(t)^T)^{-1}\dot\gamma_\ell(t)\gamma_\ell(t)^T\right) = \operatorname{ddiag}\left((BB^T)^{-1}\xi B^T\right) = 0.$$

Hence $\dot\gamma_\ell(t) \in \mathcal{H}^\ell_{\gamma_\ell(t)}$ showing that the curve $\gamma_\ell(t)$ stays horizontal in GL_n equipped with the left-invariant metric. Since $\gamma_\ell(t)$ is a complete geodesic on GL_n, it follows from proposition 2.109 in [22] that $\pi(\gamma_\ell(t))$ is a complete geodesic on $\mathcal{M}_n^{\mathrm{nh}}$. ∎

As for the oblique manifold, the extension to second-order Riemannian optimization methods on $\mathcal{M}_n^{\mathrm{nh}}$ equipped with the left-invariant metric is straightforward. Indeed, the representation of the Levi-Civita connection, used to define the Riemannian Hessian, is obtained by projecting the Levi-Civita connection of GL_n onto the horizontal space; see [2, proposition 5.3.3].

4.3. Right-invariant metric. In this section we endow GL_n with the right-invariant metric in (5). Unfortunately, this metric does not satisfy (11) and thus does not induce a well-defined Riemannian metric on the quotient manifold $\mathcal{M}_n^{\mathrm{nh}}$. Let f be an objective function on GL_n inducing a criterion $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$; we discuss here how we can still define an optimization algorithm to minimize $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$ even though the gradient of $\widehat{f}$ is not properly defined on the non-Riemannian quotient $\mathcal{M}_n^{\mathrm{nh}}$.

We start with noting that for all $B \in \mathrm{GL}_n$, we can derive the horizontal space $\mathcal{H}_B^r$ to the vertical space $\mathcal{V}_B$ according to the right-invariant metric and the associated orthogonal projection map. They are both given in proposition 4.3. Notice that $\mathcal{H}_B^r$ is also isomorphic to the tangent space $T_{\pi(B)}\mathcal{M}_n^{\mathrm{nh}}$ of $\pi(B) \in \mathcal{M}_n^{\mathrm{nh}}$ in this case.

PROPOSITION 4.3. *The horizontal space $\mathcal{H}_B^r$ at B in GL_n endowed with the right-invariant metric is*

$$\mathcal{H}_B^r = \{\xi \in \mathbb{R}^{n\times n} : \operatorname{ddiag}(\xi B^{-1}) = 0\}.$$

It follows that the orthogonal projection map at B from $\mathbb{R}^{n\times n}$ onto $\mathcal{H}_B^r$ is

$$P_B^{\mathrm{nh},r}(Z) = Z - \operatorname{ddiag}(ZB^{-1})B.$$

Proof. The set $\mathcal{H}_B^r$ is of dimension $n^2 - n$ and for all $B \in \mathrm{GL}_n$, $\Delta \in \mathcal{D}_n$ and $\xi \in \mathbb{R}^{n\times n}$, we have

$$\langle \xi, \Delta B\rangle_B^r = \operatorname{tr}(\xi B^{-1}\Delta) = \operatorname{tr}(\operatorname{ddiag}(\xi B^{-1})\Delta).$$

Thus, $\langle \xi, \Delta B\rangle_B^r = 0$ for all $\Delta \in \mathcal{D}_n$ if and only if $\operatorname{ddiag}(\xi B^{-1}) = 0$. It remains to determine the orthogonal projection map on $\mathcal{H}_B^r$. We know that $P_B^{\mathrm{nh},r}(Z) = Z - \Lambda B$ with $\Lambda \in \mathcal{D}_n$ and solving $\operatorname{ddiag}(P_B^{\mathrm{nh},r}(Z)B^{-1}) = 0$ yields the result. ∎

We want to define a *descent algorithm* in order to optimize $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$ while GL_n is equipped with the right-invariant metric. We do this by adapting the Riemannian gradient descent algorithm through the following steps:

- *Descent direction:* we exploit the Riemannian gradient $\operatorname{grad}_r f(B)$ of $f = \widehat{f}\circ\pi$ at B in GL_n. It is an element of $\mathcal{H}_B^r$ as

  $$\operatorname{ddiag}(\operatorname{grad}_r f(B)B^{-1}) = 0.$$

  Since $-\operatorname{grad}_r f(B)$ is a descent direction of f on GL_n, the corresponding tangent vector in $T_{\pi(B)}\mathcal{M}_n^{\mathrm{nh}}$ is a descent direction of $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$. Given $B \in \mathrm{GL}_n$ and $\Sigma \in \mathcal{D}_n^*$, direct computations show that the relation between the gradients of f at B and ΣB is

  $$\operatorname{grad}_r f(\Sigma B) = \Sigma^{-1}\operatorname{grad}_r f(B)B^{-1}\Sigma^2 B. \tag{13}$$

- *Failure of retraction:* the descent direction $\xi = -\operatorname{grad}_r f(B)$ in $\mathcal{H}_B^r$ at B does not correspond to the direction $\Sigma\xi$ in $\mathcal{H}_{\Sigma B}^r$. Instead, ξ corresponds to the vector $\Sigma^{-1}\xi B^{-1}\Sigma^2 B$ in $\mathcal{H}_{\Sigma B}^r$ at ΣB. As a consequence, retracting

$\xi = -\operatorname{grad}_r f(B)$ with a retraction R on GL_n satisfying (12) does not satisfy the expected property on $\mathcal{M}_n^{\mathrm{nh}}$ since

$$\pi(R_{\Sigma B}(-\operatorname{grad}_r f(\Sigma B))) \neq \pi(R_B(-\operatorname{grad}_r f(B))).$$

- *Pseudo-retraction:* in order to define a proper algorithm on $\mathcal{M}_n^{\mathrm{nh}}$, *i.e.*, which does not depend on the choice of the representative B of an equivalence class, we need to define an operator $\widetilde{R}_B : \mathcal{H}_B^r \to \mathrm{GL}_n$ such that

$$\pi(\widetilde{R}_B(\xi)) = \pi(\widetilde{R}_{\Sigma B}(\Sigma^{-1}\xi B^{-1}\Sigma^2 B)),$$

for all $B \in \mathrm{GL}_n$, $\xi \in \mathcal{H}_B^r$ and $\Sigma \in \mathcal{D}_n^*$. This property indeed ensures that the different representatives of an equivalence class in $\mathcal{M}_n^{\mathrm{nh}}$ all yield the same equivalence class. In this work, we propose to use the *pseudo-retraction* defined, for all $B \in \mathrm{GL}_n$, $\xi \in \mathcal{H}_B^r$, as

$$\widetilde{R}_B(\xi) = \exp(\Lambda(B)\xi B^{-1}\Lambda(B)^{-1} - (\xi B^{-1})^T)\exp((\xi B^{-1})^T)B,$$

where $\Lambda(B) = \operatorname{ddiag}(BB^T)$. From direct computations, we obtain that the chosen operator $\widetilde{R}$ satisfies

$$\widetilde{R}_{\Sigma B}(\Sigma^{-1}\xi B^{-1}\Sigma^2 B) = \Sigma \widetilde{R}_B(\xi),$$

showing that it possesses the adequate property. However, $\widetilde{R}$ does not define a retraction on GL_n: it satisfies $\widetilde{R}_B(0) = B$ but $\mathrm{D}\,\widetilde{R}_B(0)[\xi] \neq \xi$. Instead, we have

$$\mathrm{D}\,\widetilde{R}_B(0)[\xi] = \Lambda(B)\xi B^{-1}\Lambda(B)^{-1}B.$$

Further notice that $\widetilde{R}_B$ coincides with the Riemannian exponential map $\exp_B^r$ for B such that $\operatorname{ddiag}(BB^T) = I_n$.

- *Algorithm:* in view of the above developments, given $\mathcal{B}_i = \pi(B_i)$ in $\mathcal{M}_n^{\mathrm{nh}}$, we propose the pseudo-Riemannian gradient descent iteration of the criterion $\widehat{f}$ on $\mathcal{M}_n^{\mathrm{nh}}$ (induced by $f = \widehat{f} \circ \pi$ on GL_n) defined by

$$\mathcal{B}_{i+1} = \pi(\widetilde{R}_{B_i}(-t_i \operatorname{grad}_r f(B_i))).$$

To extend this algorithm in order to handle more sophisticated descent directions (such as the ones used in conjugate gradient or quasi-Newton methods), we need to be able to use descent directions of previous iterates. Thus, we need a *pseudo-vector transport* operator $\widetilde{T}$ on GL_n associated with the pseudo-retraction $\widetilde{R}$. This pseudo-transport must ensure that, for all the representatives of an equivalence class in $\mathcal{M}_n^{nh}$, the obtained descent directions yield the same equivalence class. Given $B \in \mathrm{GL}_n$, $\xi, \eta \in \mathcal{H}_B^r$, the purpose of $\widetilde{\mathcal{T}}(B, \xi, \eta)$ is to transport η from $\mathcal{H}_B^r$ into $\mathcal{H}_{\widetilde{R}_B(\xi)}^r$. We know from (13) that, given $\Sigma \in \mathcal{D}_n^*$, ξ and η at B correspond to $\Sigma^{-1}\xi B^{-1}\Sigma^2 B$ and $\Sigma^{-1}\eta B^{-1}\Sigma^2 B$ at ΣB. Further recall that $\widetilde{R}_{\Sigma B}(\Sigma^{-1}\xi B^{-1}\Sigma^2 B) = \Sigma \widetilde{R}_B(\xi)$. It follows that to ensure the sought invariance property, $\widetilde{\mathcal{T}}(\Sigma B, \Sigma^{-1}\xi B^{-1}\Sigma^2 B, \Sigma^{-1}\eta B^{-1}\Sigma^2 B)$ should be the vector in $\mathcal{H}_{\Sigma \widetilde{R}_B(\xi)}^r$ corresponding to $\widetilde{\mathcal{T}}(B, \xi, \eta)$ in $\mathcal{H}_{\widetilde{R}_B(\xi)}^r$. This translates into condition

$$\widetilde{\mathcal{T}}(\Sigma B, \Sigma^{-1}\xi B^{-1}\Sigma^2 B, \Sigma^{-1}\eta B^{-1}\Sigma^2 B) = \Sigma^{-1}\widetilde{\mathcal{T}}(B, \xi, \eta)\widetilde{R}_B(\xi)^{-1}\Sigma^2 \widetilde{R}_B(\xi).$$

metric	$\langle \cdot, \cdot \rangle^{\ell}$	$\langle \cdot, \cdot \rangle^{r}$
projection map	$P_B^{\mathrm{nh},\ell}(Z) = Z - \Lambda B$, $\mathrm{diag}(\Lambda) = ((BB^T)^{-1} \odot BB^T)^{-1} \mathrm{diag}(ZB^T)$	$P_B^{\mathrm{nh},r}(Z) = Z - \Lambda B$, $\Lambda = \mathrm{ddiag}(ZB^{-1})$
gradient	$\mathrm{grad}_{\ell} f(B)$	$\mathrm{grad}_{r} f(B)$
retraction	$\exp_B^{\ell}(\xi)$	$\widetilde{R}_B(\xi) = X \exp((\xi B^{-1})^T)B$, $X = \exp(\Lambda \xi B^{-1} \Lambda^{-1} - (\xi B^{-1})^T)$, $\Lambda = \mathrm{ddiag}(BB^T)$
vector transport	$P^{\mathrm{nh},\ell}_{\exp_B^{\ell}(\xi)}(\exp_B^{\ell}(\xi) B^{-1} \eta)$	$P^{\mathrm{nh},r}_{\widetilde{R}_B(\xi)}(\eta (B^TB)^{-1} \widetilde{R}_B(\xi)^T \widetilde{R}_B(\xi))$

TABLE 2
Summary of the necessary tools for optimization with the non-holonomic constraint.

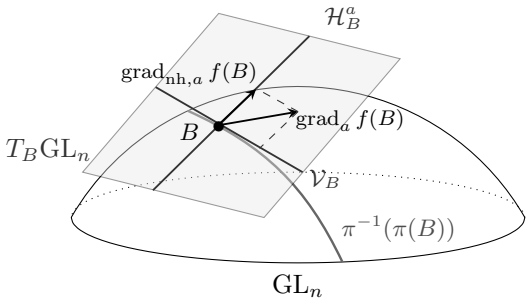

FIG. 2. *Schematic illustration of the correction on the Riemannian gradient for an unproperly defined function f on $\mathcal{M}_n^{\mathrm{nh}}$. When f is not invariant along the equivalence class $\pi^{-1}(\pi(B))$, we cancel out the action of the diagonal scaling by projecting $\mathrm{grad}_a f(B)$ onto $\mathcal{H}_B^a$. The script 'a' stands for 'r' or 'ℓ' corresponding to the right- and left-invariant metrics, respectively.*

A suitable pseudo-vector transport $\widetilde{\mathcal{T}}$ is defined for all $B \in \mathrm{GL}_n$, $\xi, \eta \in \mathcal{H}_B^r$ as

$$\widetilde{\mathcal{T}}(B, \xi, \eta) = P^{\mathrm{nh},r}_{\widetilde{R}_B(\xi)}(\eta (B^TB)^{-1} \widetilde{R}_B(\xi)^T \widetilde{R}_B(\xi)).$$

To conclude this section, a summary of all the necessary tools for optimization on $\mathcal{M}_n^{\mathrm{nh}}$ while GL_n is equipped with the left- and right-invariant metrics is given in table 2.

4.4. Unproperly defined criteria. In this section, we briefly discuss how to exploit the geometry of $\mathcal{M}_n^{\mathrm{nh}}$ to minimize an approximate joint diagonalization criterion f that is not invariant along equivalence classes, like for example the least squares criterion (8). A criterion f that changes along equivalence classes does not fully capture the approximate joint diagonalization problem, and in practice it might yield degenerate solutions when the optimization is conducted directly on GL_n. For instance, the null matrix 0, which is a limit point to every equivalence class, is a trivial solution to the optimization problem with criterion (8).

However, these mismodeled functions have some merits: for instance, (8) is simple to compute, is defined for any set $\{C_k\}$ and is widely used. To avoid degenerate solutions that such functions might produce, we take into account prior knowledge provided by the geometrical structure to compensate the mismodeling of criterion

metric	$\langle \cdot, \cdot \rangle^{\ell}$	$\langle \cdot, \cdot \rangle^{r}$
projection map	$P_B^{\mathrm{nh},\ell}(Z) = Z - \Lambda B$, $\mathrm{diag}(\Lambda) = ((BB^T)^{-1} \odot BB^T)^{-1} \mathrm{diag}(ZB^T)$	$P_B^{\mathrm{nh},r}(Z) = Z - \Lambda B$, $\Lambda = \mathrm{ddiag}(ZB^{-1})$
corrected gradient	$P_B^{\mathrm{nh},\ell}(\mathrm{grad}_\ell f(B))$	$P_B^{\mathrm{nh},r}(\mathrm{grad}_r f(B))$
retraction	$\exp_B^{\ell}(\xi)$	$\exp_B^{r}(\xi)$
corrected vector transport	$P^{\mathrm{nh},\ell}_{\exp_B^{\ell}(\xi)}(\exp_B^{\ell}(\xi)B^{-1}\eta)$	$P^{\mathrm{nh},r}_{\exp_B^{r}(\xi)}(\eta B^{-1} \exp_B^{r}(\xi))$

TABLE 3

Summary of the necessary tools for optimization of criteria that do not induce functions on $\mathcal{M}_n^{\mathrm{nh}}$ with the non-holonomic constraint.

f. Following the approach of [9], rather than changing the search space, we perform the minimization in the total manifold GL_n, however a correction is applied on the Riemannian gradient in order to cancel out the influence of the diagonal scaling.

More precisely, for the left- and right-invariant metrics in (5), the Riemannian gradients $\mathrm{grad}_\ell f(B)$ and $\mathrm{grad}_r f(B)$ of f at $B \in \mathrm{GL}_n$ both have a non zero component on the vertical space $\mathcal{V}_B$. In order to neutralize the action of the diagonal scaling, we use the technique of [5, 6, 9, 38] which consists in modifying the Riemannian gradient by canceling the part in $\mathcal{V}_B$ with the corresponding orthogonal projection. By doing so, we remove the component of the gradient that induces a displacement along the equivalence class $\pi^{-1}(\pi(B))$; see the illustration in figure 2. The corrected Riemannian gradients at $B \in \mathrm{GL}_n$ are then defined as

$$\mathrm{grad}_{\mathrm{nh},\ell} f(B) = P_B^{\mathrm{nh},\ell}(\mathrm{grad}_\ell f(B)) \qquad \text{and} \qquad \mathrm{grad}_{\mathrm{nh},r} f(B) = P_B^{\mathrm{nh},r}(\mathrm{grad}_r f(B)).$$

Similarly, we also apply a correction on the vector transport to cancel the action of the diagonal scaling. The corrected vector transports are given by

$$\mathcal{T}_{\mathrm{nh},\ell}(B,\xi,\eta) = P^{\mathrm{nh},\ell}_{\exp_B^{\ell}(\xi)}(\exp_B^{\ell}(\xi)B^{-1}\eta)$$
$$\text{and} \qquad \mathcal{T}_{\mathrm{nh},r}(B,\xi,\eta) = P^{\mathrm{nh},r}_{\exp_B^{r}(\xi)}(\eta B^{-1} \exp_B^{r}(\xi)).$$

Note that the corrected Riemannian gradients are still not invariant to the choice of the representant B of an equivalence class $\pi^{-1}(\pi(B))$ in the general case. Thus, unlike in the previous sections, the optimization algorithms using these corrected gradients are not defined on the quotient $\mathcal{M}_n^{\mathrm{nh}}$ but on GL_n and depend on the chosen representative of an equivalence class.

We conclude this section with a summary of all the necessary tools for optimization of criteria that do not induce functions on $\mathcal{M}_n^{\mathrm{nh}}$ with the non-holonomic constraint in table 3.

5. Numerical illustrations. We conduct numerical experiments to illustrate the applicability and versatility of our theoretical results. We construct AJD methods by optimizing the criteria given in section 2.3 using the tools developped in sections 3 and 4. We first study the performance of AJD methods on simulated symmetric positive definite matrices in section 5.1 and we then look at the results obtained on an example of real electroencephalographic (EEG) data in section 5.2.

$I_{\text{M-A}}$ (dB)	ffdiag	jadiag	F-nh-r	$\widetilde{\text{F}}$	ll
$\sigma = 100$	-17.75 ± 1.92	-16.07 ± 1.28	-17.09 ± 1.86	-17.93 ± 1.89	-16.07 ± 1.28
$\sigma = 1000$	-23.16 ± 3.75	-19.82 ± 2.69	-22.44 ± 4.06	-23.45 ± 3.66	-19.82 ± 2.69

TABLE 4

Mean and standard deviation of the Moreau-Amari index over 500 trials of algorithms ffdiag [38] and jadiag [30] and of the methods developed in this paper that gave the best results on these simulated data. $\widetilde{\text{F}}$ *and ll correspond to the results obtained with all algorithms based on criteria* (9) *and* (10) *respectively since they are identical once they have converged.*

We denote the algorithms as follows: we first indicate the optimized cost function with 'F', '$\widetilde{\text{F}}$' or 'll' corresponding to (8), (9) or (10), respectively; we then indicate the chosen constraint with 'ob' or 'nh' for the oblique and non-holonomic constraint; finally we indicate whether we use the left- or right-invariant metrics with 'ℓ' or 'r'. Optimization is performed with the Riemannian BFGS method proposed in [23] and implemented in the manopt toolbox [12]. The stopping criterion for iterate B_i is defined as $\|B_{i-1}^{-1}B_i - I_n\|_F^2/n$ and its tolerance is set to 10^{-12}.

5.1. Simulated data. We simulate sets of $K = 50$ real valued $n \times n$ (with $n = 32$) symmetric positive definite matrices C_k according to model

$$C_k = A\Lambda_k A^T + \frac{1}{\sigma} E_k \Delta_k E_k^T, \tag{14}$$

where matrices A and E_k are random matrices with independent and identically distributed (i.i.d.) elements drawn from the standard normal distribution. Diagonal matrices Λ_k and Δ_k simulate signal and noise source energies and have i.i.d. elements drawn from the chi-squared distribution with expectation 1. Free parameter σ defines the expected signal to noise ratio of the data. In our experiments we consider $\sigma = 1000$, which corresponds to a very good signal to noise ratio, and $\sigma = 100$, for which matrices are further from being jointly diagonalizable but a satisfying estimation of A can still be obtained. Before performing AJD, we do a pre-whitening of matrices C_k with the inverse square root of their arithmetic mean and initialize all algorithms with the identity matrix.

We estimate the performance of algorithms by using a standard measure of accuracy for AJD (the so-called Moreau-Amari index [29])

$$I_{\text{M-A}}(M) = \frac{1}{2n(n-1)} \sum_{p=1}^{n} \left(\frac{\sum_{q=1}^{n} |M_{pq}|}{\max\limits_{1 \le q \le n} |M_{pq}|} + \frac{\sum_{q=1}^{n} |M_{qp}|}{\max\limits_{1 \le q \le n} |M_{qp}|} - 2 \right),$$

where $M = BA$, with B the estimated joint diagonalizer and A the true mixing matrix of the signal part in (14). Thus, $I_{\text{M-A}}$ is a measure in $[0, 1]$ with zero indicating a perfect recovering of the mixing process. In the following, the Moreau-Amari index is reported in decibel (dB) computed with the formula $10\log(I_{\text{M-A}})$, thus, the lower the index, the better.

We first compare the performance in terms of accuracy of the proposed algorithms with two previously published ones: ffdiag [38], which minimizes the Frobenius criterion (8) and exploits the non-holonomic constraint, and jadiag [30], which minimizes the log-likelihood criterion (10). As shown in table 4, the results of the algorithms

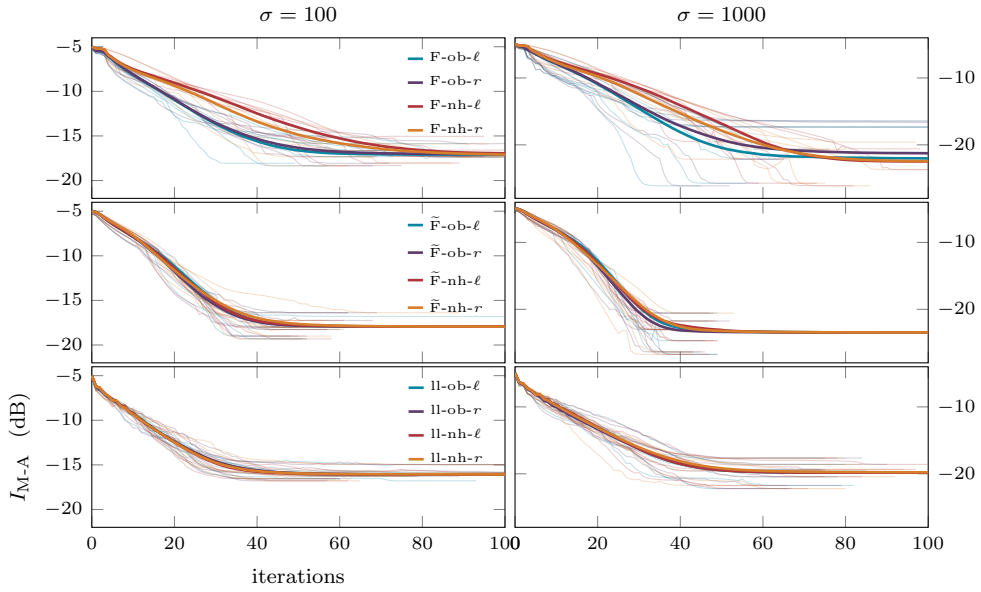

FIG. 3. *Mean over 500 trials (dark) along with a few single trials (faded) of the Moreau-Amari index obtained with proposed AJD algorithms on the simulated data following model* (14) *as a function of the number of iterations. The signal to noise ratio parameter* σ *is* 100 *on the left and* 1000 *on the right. From top to bottom, these results correspond to algorithms optimizing the Frobenius, modified Frobenius and log-likelihood criteria* (8), (9) *and* (10) *respectively.*

we have proposed based on the log-likelihood are identical to those obtained with jadiag. Moreover, the performance of ffdiag is close to the one of proposed methods exploiting the non-holonomic constraint and minimizing the Frobenius criterion and of those optimizing the modified Frobenius criterion. These simulations show that the generic geometrical approach developed in this work reaches the state of the art of specific algorithms.

As illustrated in figure 3, for $\sigma = 1000$, F-ob-r often converges toward a good solution, but it fails in 12% of cases. F-ob-ℓ also misconverges in 4.2% of cases. It thus appears that, on these simulated data, the left-invariant metric is more advantageous than the right-invariant metric for the Frobenius criterion (8) when the oblique constraint is employed. On the other hand, the performance of the algorithms based on the non-holonomic constraint is consistent. Their convergence in terms of iterations is, however, usually slower. Thus, the correction applied to the descent direction presented in section 4.4 allows to obtain the best results with the Frobenius criterion (8) on these data. The final results obtained with the methods minimizing the modified Frobenius and log-likelihood criteria (9) and (10) do not depend on the choice of the metric and constraint. On average, the convergence of the different algorithms is quite similar for both criteria.

5.2. Real electroencephalographic data. We test our AJD algorithms by performing the BSS of an electroencephalographic (EEG) recording[1] where the subject

[1]Data taken from subject 11 of a dataset concerning an experiment using the brain-computer interface [17]. Signals were acquired with 16 electrodes and sampled at 512Hz. For analysis, they were filtered in the band-pass 1-20Hz and down-sampled to 128Hz. Full details on data and preprocessing can be found in [18].

has to focus on a particular target among a set of displayed objects that randomly flash in predefined sequences. These data comprise two classes (target and non-target) depending on whether a given flash concerns the target or not. Such experiment is a typical visual odd-ball paradigm where event-related potentials are evoked. To perform the blind source separation of a set of mixed signals through AJD, the choice of the matrices to be jointly diagonalized is crucial and significantly influences the results; see *e.g.* [15] for an overview. In this work, we construct the set of matrices following the method proposed in [20], which is adapted to the analysis of EEG signals containing event-related potentials. We take the two covariance matrices of the means of the signals of the two classes along with the 20 Fourier cospectra between 1-20Hz with 1Hz resolution. We thus have a set of 22 matrices of size 16×16 to diagonalize.

Compared to the previous example on simulated data, there is no ground truth to validate and compare the obtained results. First, we perform a general pair-wise comparison by means of the index

$$\widetilde{I}_{\text{M-A}} = (I_{\text{M-A}}(B\widehat{B}^{-1}) + I_{\text{M-A}}(\widehat{B}B^{-1}))/2,$$

which estimates the similarity between the results B and $\widehat{B}$ of two methods. $\widetilde{I}_{\text{M-A}}$ is a measure in $[0, 1]$ with zero indicating equivalent solutions. The three following groups contain methods that yielded identical within-group results ($\widetilde{I}_{\text{M-A}} < -50$ dB in all pair-wise comparisons): *(i)* F-ob-ℓ and F-ob-r; *(ii)* the four algorithms based on the modified Frobenius criterion (9); *(iii)* the four algorithms based on the log-likelihood function (10). The other results show that the indexes of F-nh-ℓ compared to F-ob-ℓ and F-ob-r is -18.5 dB in both cases, while $\widetilde{I}_{\text{M-A}}$ is comprised between -15.7 and -14.7 dB for all remaining pair-wise comparisons of algorithms based on the Frobenius and modified Frobenius criteria (8) and (9). Finally, the indexes $\widetilde{I}_{\text{M-A}}$ between the four log-likelihood algorithms and the others are in the range $[-8.3, -8]$ dB. We can thus distinguish five groups of solutions with this index: F-ob-ℓ and F-ob-r; F-nh-ℓ; F-nh-r; the four modified Frobenius algorithms; the four log-likelihood algorithms.

Then, we examine the event-related potential sources obtained with the different methods, which are shown in figure 4. Source s1, which is retrieved by all algorithms, appears to be the most discriminative source between the two classes. It explains most of the variance in the mean event-related potential of the target class and is located near the visual cortex. Source s2, only found by methods based on the log-likelihood criterion, is located around central electrode Cz and active in both conditions. Its localization and its mean in the non-target condition are similar to those of source s4, which is found by algorithms based on the modified Frobenius criterion and F-nh-r. Source s4 seems to be decomposed into the two sources, s5 and s6, found by methods F-ob-ℓ, F-ob-r and F-nh-ℓ. Finally, another plausible event-related potential source appears to be s3, which is more frontal and obtained with all algorithms based on the Frobenius and modified Frobenius criteria. We thus have obtained three groups of solutions: those obtained by F-ob-ℓ, F-ob-r and F-nh-ℓ (s1,s3,s5,s6); those obtained by F-nh-r and algorithms based on the modified Frobenius criterion (s1,s3,s4); and those obtained by methods based on the log-likelihood (s1,s2).

5.3. Summary on numerical results. These preliminary numerical illustrations lead us to several observations concerning our optimization framework for AJD. First, solutions obtained with the modified Frobenius and log-likelihood criteria, which share the diagonal scaling invariance property, appear not to be influenced by the choice of the metric and constraint both on simulated and real EEG data. Simulated data showed that the choice of the constraint can influence the optimization of the

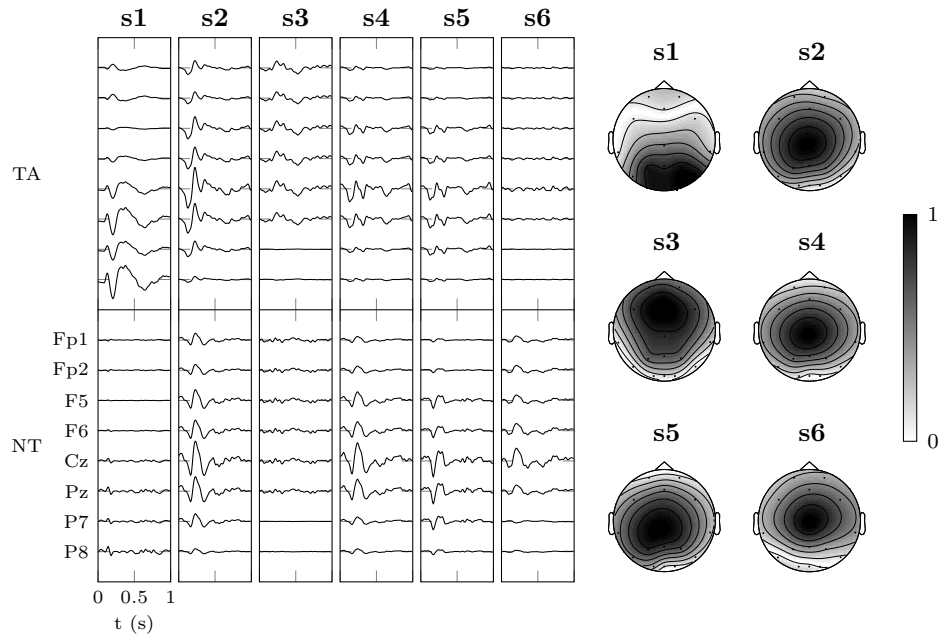

FIG. 4. *Projection of the mean for each class (target - TA and non-target - NT) of six plausible event-related potential sources retrieved by the algorithms on selected electrodes alongside their normalized spatial distribution on the scalp. Source s1 is obtained with all algorithms; s2 is retrieved with algorithms based on the log-likelihood; s3 is obtained with all methods relying on the Frobenius and modified Frobenius criteria; s4 comes from algorithms optimizing the modified Frobenius criterion and F-nh-r; s5 and s6 are retrieved with algorithms F-ob-ℓ, F-ob-r and F-nh-ℓ. We checked that the values of spatial and temporal correlations of sources commonly obtained by different methods are above 0.99.*

Frobenius criterion and that the non-holonomic constraint seems more advantageous in this case. On real data, the choice of the metric yields different results for the Frobenius criterion with the non-holonomic constraint. However, these results may be specific to the considered data and deserve further study before coming to conclusions on these applications. This is beyond the scope of this paper, which aim is to present an unified optimization approach along with all necessary geometrical ingredients.

6. Conclusions and perspectives. In this work we have developed a unified Riemannian optimization framework for AJD constructed directly from the general linear group and handling the two standard oblique and non-holonomic constraints. As our formulation is general, it can be used with any AJD objective function (providing only its Euclidean gradient) and with a large panel of Riemannian gradient-based optimization algorithms.

We have endowed the general linear group with two metrics (left- and right-invariant), which have opposite interests and drawbacks. The left-invariant metric, whose use is original in this context, suits well the structure of the AJD solution and enabled us to define a proper Riemannian quotient manifold for the non-holonomic constraint. On the other hand, the right-invariant metric has a natural link with the structure of the AJD model, but leads to a non-Riemannian quotient manifold for the non-holonomic constraint. We formalized previous works and developed new

tools beyond Riemannian optimization to obtain algorithms properly defined on the quotient.

This work opens perspectives of research, including mathematical questions on optimization in this context (such as geodesical convexity of the considered AJD criteria) and application or generalization to related problems (*e.g.*, independent vector analysis [24], joint independent subspace analysis [25], bilinear BSS [19]).

REFERENCES

[1] P.-A. Absil and K. A. Gallivan, *Joint diagonalization on the oblique manifold for independent component analysis*, in Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, vol. 5, May 2006, pp. 945–948.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, USA, 2008.

[3] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.

[4] B. Afsari, *Simple LU and QR based non-orthogonal matrix joint diagonalization*, in International Conference on Independent Component Analysis and Signal Separation, Springer, 2006, pp. 1–7.

[5] B. Afsari, *Sensitivity analysis for the problem of matrix joint diagonalization*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 1148–1171.

[6] B. Afsari and P. S. Krishnaprasad, *Some gradient based joint diagonalization methods for ICA*, in Independent Component Analysis and Blind Signal Separation, Springer, 2004, pp. 437–444.

[7] K. Alyani, M. Congedo, and M. Moakher, *Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem*, Linear Algebra and its Applications, (2016).

[8] S.-I. Amari, *Natural gradient works efficiently in learning*, Neural computation, 10 (1998), pp. 251–276.

[9] S.-I. Amari, T. Chen, and A. Cichocki, *Nonholonomic orthogonal learning algorithms for blind source separation*, Neural computation, 12 (2000), pp. 1463–1484.

[10] E. Andruchow, G. Larotonda, L. Recht, and A. Varela, *The left invariant metric in the general linear group*, Journal of Geometry and Physics, 86 (2014), pp. 241–257.

[11] F. Bouchard, J. Malick, and M. Congedo, *Riemannian optimization and approximate joint diagonalization for blind source separation*, IEEE Transactions on Signal Processing, 66 (2018), pp. 2041–2054.

[12] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt, a Matlab toolbox for optimization on manifolds*, Journal of Machine Learning Research, 15 (2014), pp. 1455–1459.

[13] J.-F. Cardoso and A. Souloumiac, *Blind beamforming for non Gaussian signals*, IEEE Proceedings-F, 140 (1993), pp. 362–370.

[14] J.-F. Cardoso and A. Souloumiac, *Jacobi angles for simultaneous diagonalization*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 161–164.

[15] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavsky, and A. Yeredor, *Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications*, IEEE Signal Processing Magazine, 31 (2014), pp. 34–43.

[16] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 1st ed., 2010.

[17] M. Congedo, M. Goyat, N. Tarrin, G. Ionescu, L. Varnet, B. Rivet, R. Phlypo, N. Jrad, M. Acquadro, and C. Jutten, *"Brain Invaders": a prototype of an open-source p300-based video game working with the openvibe platform*, in 5th International Brain-Computer Interface Conference 2011 (BCI 2011), 2011, pp. 280–283.

[18] M. Congedo, L. Korczowski, A. Delorme, and F. Lopes da Silva, *Spatio-temporal common pattern: A companion method for ERP analysis in the time domain*, Journal of neuroscience methods, 267 (2016), pp. 74–88.

[19] M. Congedo, R. Phlypo, and D. T. Pham, *Approximate joint singular value decomposition of an asymmetric rectangular matrix set*, IEEE Transactions on Signal Processing, 59 (2011), pp. 415–424.

[20] M. Congedo, S. Rousseau, and C. Jutten, *An introduction to EEG source analysis with an illustration of a study on error-related potentials*, in Guide to Brain-Computer Music

Interfacing, Springer, 2014, pp. 163–189.
[21] B. N. Flury and W. Gautschi, *An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 169–184.
[22] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry*, Springer, 3rd ed., 2004.
[23] W. Huang, P.-A. Absil, and K. Gallivan, *A Riemannian BFGS Method for Nonconvex Optimization Problems*, Springer International Publishing, 2016, pp. 627–634.
[24] T. Kim, T. Eltoft, and T. Lee, *Independent vector analysis: An extension of ica to multivariate components*, in International Conference on Independent Component Analysis and Signal Separation, Springer, 2006, pp. 165–172.
[25] D. Lahat and C. Jutten, *Joint independent subspace analysis using second-order statistics*, IEEE Transactions on Signal Processing, 64 (2016), pp. 4891–4904.
[26] J. Lee, *Introduction to Smooth Manifolds*, Graduate Texts in Mathematics, Springer, 2003.
[27] S. Lee, M. Choi, H. Kim, and F. C. Park, *Geometric direct search algorithms for image registration*, IEEE Transactions on Image Processing, 16 (2007), pp. 2215–2224.
[28] M. I. Miller, A. Trouvé, and L. Younes, *The metric spaces, Euler equations, and normal geodesic image motions of computational anatomy*, in Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, vol. 2, IEEE, 2003, pp. 635–638.
[29] E. Moreau and O. Macchi, *A one stage self-adaptive algorithm for source separation*, in IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., vol. 3, Apr 1994, pp. 49–52.
[30] D.-T. Pham, *Joint approximate diagonalization of positive definite Hermitian matrices*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 1136–1152.
[31] D.-T. Pham and J.-F. Cardoso, *Blind separation of instantaneous mixtures of nonstationary sources*, IEEE Transactions on Signal Processing, 49 (2001), pp. 1837–1848.
[32] D.-T. Pham and M. Congedo, *Least square joint diagonalization of matrices under an intrinsic scale constraint*, in Independent Component Analysis and Signal Separation, Springer, 2009, pp. 298–305.
[33] A. Souloumiac, *Nonorthogonal joint diagonalization by combining Givens and hyperbolic rotations*, Signal Processing, IEEE Transactions on, 57 (2009), pp. 2222–2231.
[34] P. Tichavskỳ and A. Yeredor, *Fast approximate joint diagonalization incorporating weight matrices*, Signal Processing, IEEE Transactions on, 57 (2009), pp. 878–891.
[35] B. Vandereycken, P.-A. Absil, and S. Vandewalle, *A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank*, IMA Journal of Numerical Analysis, 33 (2012), pp. 481–514.
[36] A. Yeredor, A. Ziehe, and K.-R. Müller, *Approximate joint diagonalization using a natural gradient approach*, in Independent Component Analysis and Blind Signal Separation, Springer, 2004, pp. 89–96.
[37] E. Zacur, M. Bossa, and S. Olmos, *Left-invariant Riemannian geodesics on spatial transformation groups*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1503–1557.
[38] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, *A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation*, The Journal of Machine Learning Research, 5 (2004), pp. 777–800.